中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2023)09-2817-16

论文引用格式:Si Z F and Qi H G. 2023. Survey on knowledge distillation and its application. Journal of Image and Graphics, 28(09):2817-2832(司兆峰, 齐洪钢. 2023. 知识蒸馏方法研究与应用综述. 中国图象图形学报, 28(09):2817-2832)[DOI:10.11834/jig.220273]

知识蒸馏方法研究与应用综述

司兆峰,齐洪钢*

中国科学院大学计算机科学与技术学院,北京　100049

摘　要:随着深度学习方法的不断发展,其存储代价和计算代价也不断增长,在资源受限的平台上,这种情况给其应用带来了挑战。为了应对这种挑战,研究者提出了一系列神经网络压缩方法,其中知识蒸馏是一种简单而有效的方法,成为研究热点之一。知识蒸馏的特点在于它采用了"教师—学生"架构,使用一个大型网络指导小型网络进行训练,以提升小型网络在应用场景下的性能,从而间接达到网络压缩的目的。同时,知识蒸馏具有不改变网络结构的特性,从而具有较好的可扩展性。本文首先介绍知识蒸馏的由来以及发展,随后根据方法优化的目标将知识蒸馏的改进方法分为两大类,即面向网络性能的知识蒸馏和面向网络压缩的知识蒸馏,并对经典方法和最新方法进行系统的分析和总结,最后列举知识蒸馏方法的几种典型应用场景,以便加深对各类知识蒸馏方法原理及其应用的理解。知识蒸馏方法发展至今虽然已经取得较好的效果,但是各类知识蒸馏方法仍然有不足之处,本文也对不同知识蒸馏方法的缺陷进行了总结,并根据网络性能和网络压缩两个方面的分析,给出对知识蒸馏研究的总结和展望。

关键词:知识蒸馏;深度学习;计算机视觉;神经网络;模型压缩

Survey on knowledge distillation and its application

Si Zhaofeng, Qi Honggang*

School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing　100049, China

Abstract: Deep learning is an effective method in various tasks, including image classification, object detection, and semantic segmentation. Various architectures of deep neural networks (DNNs), such as Visual Geometry Group network (VGGNet), residual network(ResNet), and GoogLeNet, have been proposed recently. All of which have high computational costs and storage costs. The effectiveness of DNNs mainly comes from their high capacity and architectural complexity, which allow them to learn sufficient knowledge from datasets and generalize well to real-world scenes. However, high capacity and architectural complexity can also result in a drastic increase in storage and computational costs, thereby complicating the implementation of deep learning methods on devices with limited resources. Given the increasing demand for deep learning methods on portable devices, such as mobile phones, the cost of DNNs must be urgently reduced. Researchers have developed a series of methods called model compression to solve the aforementioned problem. These methods can be divided into four main categories: network pruning, weight quantization, weight decomposition, and knowledge distillation. Knowledge distillation is a comparably new method first introduced in 2014. It attempts to transfer the knowledge learned by a cumbersome network (teacher network) to a lightweight network (student network), thereby allowing the student network to perform similarly to the teacher network. Thus, compression can be achieved by using the student network

收稿日期:2022-03-28;**修回日期**:2022-07-22;**预印本日期**:2022-07-29

*通信作者:齐洪钢　hgqi@ucas.ac.cn

基金项目:国家自然科学基金项目(62271466)

Supported by:National Natural Sciences Foundation of China (62271466)

for inference. Traditional knowledge distillation works by providing softened labels to the student network as the training target instead of allowing the student network to learn ground truth directly. The student network can learn about the correlation among classes in the classification problem by learning from softened labels. This approach can be taken as extra supervision while training. The student network trained by knowledge distillation should ideally approximate the performance of the teacher network. In this way, the computational and storage costs in the compressed network are reduced with minor degradation compared with those in the uncompressed network. However, this situation is almost unreachable when the compression rate is large enough to be comparable with the compression rates of other model compression methods. On the contrary, knowledge distillation can be taken as a measure of enhancing the performance of a deep learning model. Thus, this model can perform better than other models of similar size. Moreover, knowledge distillation is a method of model compression. In this study, we aim to review the knowledge distillation methods developed in recent years from a new perspective. We sort the existing methods according to their target by dividing them into performance-oriented methods and compression-oriented methods. Performance-oriented methods emphasize the improvement of the performance of the student network, whereas compression-oriented methods focus on the relationship between the size of the student network and its performance. We further divide these two categories into specific ideas. In performance-oriented methods, we describe state-of-the-art methods in two aspects: the representation of knowledge and ways of learning knowledge. The representation of knowledge has been widely studied in recent years. The researchers attempt to derive knowledge from the teacher network instead of outputting vectors to enrich the knowledge while training. Other forms of knowledge include a middle-layer feature map, representation extracted from the middle layer, and structural knowledge. The student network can learn about the teacher network's behavior while forward propagating by combining this extra knowledge with the soft target in traditional knowledge distillation. Thus, the student network acts similarly to the teacher network. Studies on the way of learning knowledge attempt to explore distillation architectures on the basis of the teacher-student architecture. Moreover, architectures includingonline distillation, self-distillation, multiteacher distillation, progressive knowledge distillation, and generative adversarial network (GAN)-based knowledge distillation are proposed. These architectures focus on the effectiveness of distillation and different use cases. For example, online distillation and self-distillation can be applied when the teacher network with high capacity is unavailable. In compression-oriented knowledge distillation, researchers try to combine neural architecture search (NAS) methods with knowledge distillation to balance the relationship between the performance and the size of the student network. Many studies on the impact of the size difference between the teacher network and the student network on distillation performance are also available. They concluded that a wide gap between the teacher and the student can cause performance degradation. Then, bridging the gap between the teacher and the student with several middle-sized networks was proposed in these studies. We also formalize different kinds of knowledge distillation methods. The corresponding figures are shown uniformly to help researchers understand the basic ideas comprehensively and learn about recent works on knowledge distillation. One of the most notable characteristics of knowledge distillation is that the architectures of the teacher network and the student network stay intact during training. Thus, other methods for different tasks can be incorporated easily. In this study, we introduce recent works on different knowledge distillation tasks, including object detection, face recognition, and natural language processing. Finally, we summarize the knowledge distillation methods mentioned before and propose several possible ideas. Recent research on knowledge distillation has mainly focused on enhancing the performance of the student network. The major problem of the student network lies in finding a feasible source of knowledge from the teacher network. Moreover, compression-based knowledge distillation suffers from the problem of searching space when NAS adjusts network architecture. On the basis of the analysis above, we propose three possible ideas for researchers to study: 1) obtaining knowledge from various tasks and architectures in the form of knowledge distillation, 2) developing a searching space for NAS when combined with knowledge distillation and adjusting the teacher network while searching for the student network, and 3) developing a metric for knowledge distillation and other model compression methods to evaluate both task performance and compression performance.

Key words: knowledge distillation; deep learning; computer vision; neural network; model compression

0 引 言

随着硬件计算能力的不断发展，深度学习方法（LeCun 等，2015）在目标检测、语义分割和自然语言处理等任务中取得越来越好的成果。为了使深度神经网络获得更好的性能，研究人员对神经网络的结构进行了不断改进，其中包括 VGGNet（Visual Geometry Group network）（Simonyan 和 Zisserman，2014）、ResNet（residual network）（He 等，2016）和 GoogLeNet（Szegedy 等，2015）等经典结构。这些网络之所以有很好的性能，一个很重要的因素在于网络结构的复杂性使得网络能够从数据集中学习到足够的知识，并将这些知识泛化到实际场景中。然而，网络结构复杂度提升的同时也带来了极高的运算代价和存储代价，以至于深度学习模型难以部署到资源受限的平台上。为了解决这个问题，神经网络压缩方法进入了研究者的视线。

神经网络压缩方法可以分为 4 类，即网络剪枝（network pruning）、权重量化（weight quantization）、权重分解（weight decomposition）和知识蒸馏（knowledge distillation，KD）。网络剪枝是其中最直接的方法，它考虑网络结构中存在的冗余连接，通过直接除去这些冗余连接来达到压缩效果。权重量化是从存储权重的角度考虑，使用低精度数值代替高精度数值从而减小存储代价。权重分解方法从运算的角度入手，简化网络运算以减小网络的运算代价。而知识蒸馏方法则是考虑神经网络整体，用学习的方法借助大型网络指导训练一个与其性能相当的小型网络，从而间接达到压缩的目的。由于这种较为抽象的提取知识的方式不会更改原本网络结构，知识蒸馏具有更好的扩展性，能够与其他网络压缩方法同时使用，而且也能方便地应用于其他任务的训练中。

知识蒸馏最先由 Hinton 等人（2015）提出，用于图像分类任务，其核心是“教师—学生”架构，即使用一个大型网络（称为教师网络）指导小型网络（称为学生网络）进行训练，以提升学生网络的性能。这种指导最早来自于教师网络输出的概率向量，学生网络通过将其作为学习目标，以获取教师网络从数据集中学习到的类间关系，这也就是最早对知识蒸馏中“知识”的定义。而“蒸馏”则来自于温度参数，它作用于产生概率向量的 softmax 层，使得类间关系更加易于学习。在知识蒸馏的训练过程中，温度参数被设置为较大的值，而当实际应用学生网络时则不使用温度参数，这就是知识蒸馏的含义。

当前，对知识蒸馏的研究主要集中在研究“知识”上，包括知识的表示形式以及知识的学习方式。对知识表示形式的探索以 Romero 等人（2015）提出的 FitNets（fit networks）为开端，尝试使用包括中间层信息、类间信息等形式来丰富学生网络从教师网络处学习到的知识，以达到提升学生网络性能的目的。另一方面，研究者尝试对“教师—学生”架构进行优化，提出了包括在线蒸馏、自蒸馏和多教师蒸馏在内的多种不同的学习结构以提高知识蒸馏的效率，它们为不同场景下的应用提供了更多的选择。

虽然知识蒸馏是作为网络压缩方法提出的，但是近年来的研究大多关注于提升学生网络的性能，于是衍生出许多用于提升网络性能的应用。与此同时，还有一些工作关注于神经网络压缩的效率，如基于神经架构搜索（neural architecture search，NAS）的知识蒸馏，在使用“教师—学生”架构的基础上，探索一个合适的网络结构以权衡网络性能与压缩率。除此之外，研究者对“教师—学生”结构中网络规模的关系进行探索，并提出了能够将知识蒸馏应用在更大压缩率的场景中的方法。

Wang 等人（2020a）和 Gou 等人（2021）对知识蒸馏进行了较为完整的综述，其中，Wang 等人（2020a）从“教师—学生”架构方面进行了总结与展望，而 Gou 等人（2021）则是从知识的表示形式的角度出发，对不同形式的知识以及学习方式进行了较为全面的综述。与先前的工作不同，本文从一个全新的视角，将近年来的知识蒸馏方法按照优化的目标分为两类，即面向网络性能的知识蒸馏和面向网络压缩的知识蒸馏，详细介绍了不同目标下知识蒸馏的发展和最新成果。本文对知识蒸馏的分类方式如图 1 所示。

1 传统知识蒸馏

最早的知识蒸馏方法（Hinton 等，2015）采用最直接的知识传输形式，即让学生网络直接学习教师网络的输出。这一过程类似于人类课堂学习，学生在教师指导下进行学习，能够使得学习的过程更加

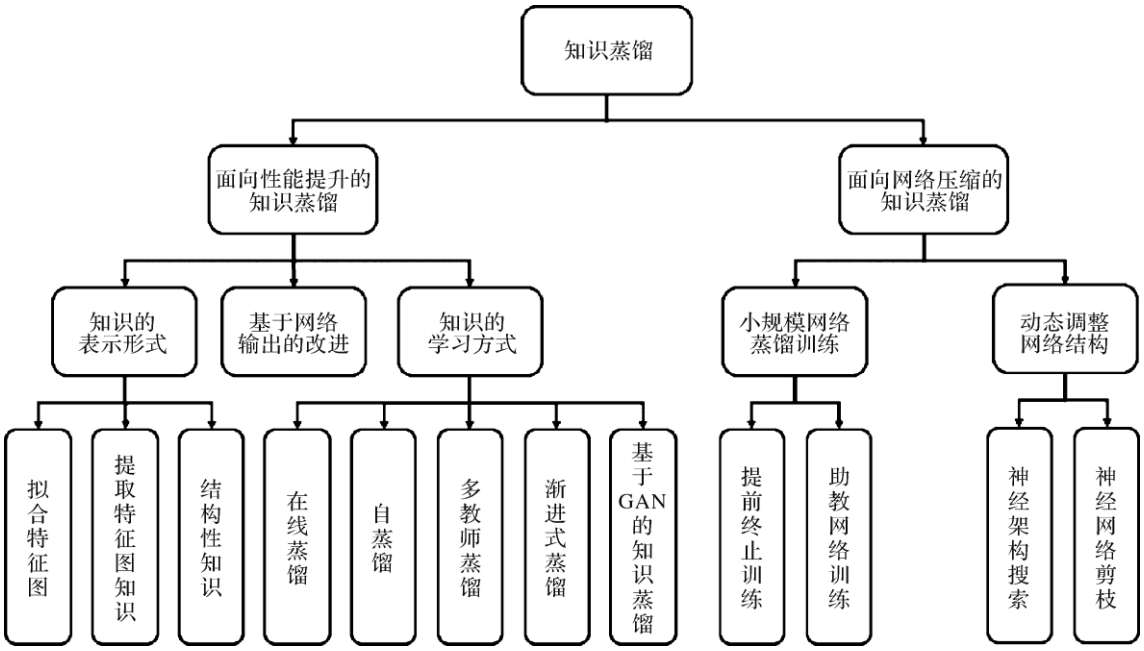

图1　知识蒸馏的分类

Fig. 1　Categories of knowledge distillation

高效。与人类教学情况类似，知识蒸馏能够帮助学生网络达到比自身直接进行学习更好的效果。传统的知识蒸馏遵循“教师—学生”架构，其损失函数可以表示为

$$L = \lambda L_{\text{task}}(\boldsymbol{l}_{\text{S}}, \boldsymbol{t}) + (1-\lambda)T^2 L_{\text{KD}}(\boldsymbol{l}_{\text{S}}, \boldsymbol{l}_{\text{T}}) \tag{1}$$

式中，L_{task}表示学生网络学习原任务的损失，也就是“课本”，$\boldsymbol{l}_{\text{s}}$是学生网络的输出向量，$\boldsymbol{t}$表示任务目标，$L_{\text{KD}}$用来计算教师网络与学生网络输出之间的差距，相当于“讲课”，$\boldsymbol{l}_{\text{T}}$是教师网络的输出向量，$\lambda$是平衡二者的系数，$T$是温度参数，用来软化标签，使其分布更为均衡。在Hinton等人(2015)的工作中，L_{KD}使用KL(Kullback-Leibler)散度作为度量方式，用以衡量二者分布的相似程度。传统的知识蒸馏形式如图2所示。

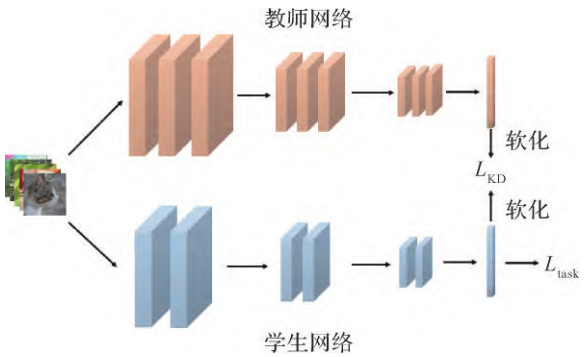

图2　传统知识蒸馏

Fig. 2　Traditional knowledge distillation

2　面向网络性能的知识蒸馏

知识蒸馏的本质是对小型网络的一种增强方法。从这个角度看，知识蒸馏方法的提升既可以是同等性能损失下压缩率的提升，也可以是同等压缩率下性能损失的减小，其中对后者的研究较多。对于提升学生网络的性能，研究者对传统知识蒸馏基于网络输出的策略进行了探究，同时在额外的监督信息方面，从丰富知识的形式以及改善学习的方法两个方面展开了研究。

2.1　基于网络输出的改进

对传统知识蒸馏的改进，即基于网络输出的改进主要关注于寻找一个更好的损失函数。Meng等人(2019)认为学生网络在学习过程中应该判断教师网络给出的输出是否可靠，而判断依据就是教师网络是否正确分类，若教师网络的知识是错误的，那么学生网络将直接学习真实目标。类似地，选择性学习的思路也用于Zhang等人(2020b)的工作，不同于选择教师网络的输出，该方法尝试对训练样本的重要程度进行分析并进行加权，将教师网络输出和学生网络输出的差距建模为高斯分布，根据学习到的标准差对样本进行加权。

Oki等人(2020)从损失函数形式的角度对传统知识蒸馏进行优化。借鉴度量学习中三元损失的思

路，尝试使用教师网络的输出作为基准，使学生网络对应样本的输出向量与之接近，并使其他样本对应的输出向量与之远离。Tian 等人（2022）则从对比学习的思路出发，尝试最大化教师网络和学生网络互信息的下界，并将其作为损失函数进行优化。

对于温度参数，Wen 等人（2021）提出了根据样本的不同在训练时使用不同温度参数的方法，采用焦点损失（focal loss）的思路和正则化概率向量的思路计算置信度，并据此计算样本的置信度以动态调整温度参数。

考虑到知识蒸馏中教师网络的计算量，Xu 等人（2020）尝试使用减少教师网络计算次数的方式降低知识蒸馏训练的代价，根据学生网络的输出计算样本的不确定性，并将不确定性大的样本进行混合用于蒸馏训练，以较小的代价获得不低于知识蒸馏的性能。

传统基于网络输出的知识蒸馏也存在一定局限性。首先，神经网络的输出只包含少量信息，蕴藏在网络结构中的大量知识仍然没有得到有效利用。同时，学习神经网络的输出是对其包含的知识的隐式学习，可解释性不强，难以使学生网络学到更加具体的知识。

2.2　探索知识的表示形式

知识蒸馏的实质是教师网络将自己学习到的知识传输给学生网络，于是“什么是知识”这一问题成为研究重点之一。传统知识蒸馏使用教师网络的输出作为知识来源，这种知识形式存在一个显著的缺陷，即包含的信息量十分有限。对于知识蒸馏而言，除了网络输出中隐式包含的有限信息，还有更多信息可以从神经网络内部提取出来。

2.2.1　直接拟合特征图

从神经网络内部提取信息相当于让学生网络学习教师网络“思考”的过程。对于这个思路，最直接的想法就是选取学生网络和教师网络中对应的层进行匹配，使得学生网络前向传播的过程接近教师网络，其损失函数可以表示为

$$L_{KD} = DF(f_S, f_T) \tag{2}$$

式中，f_S 和 f_T 分别是从学生网络和教师网络中间层提取的特征图，$DF(\cdot)$ 是距离函数（distance function，DF），如 L2 范数等。直接拟合特征图的知识蒸馏过程如图 3 所示。这一想法最早由 Romero 等人（2015）实现。他们在知识蒸馏过程中首先引入提示损失（hint loss），使用提示层（hint layer）对教师网络的特征图进行维度统一化，并将其与学生网络特征图的 L2 距离作为损失。Gao 等人（2019）简单地对这个过程进行改进，采用渐进的形式，让学生网络以提示层的形式逐层学习教师网络的特征图，以获得更为全面的监督。

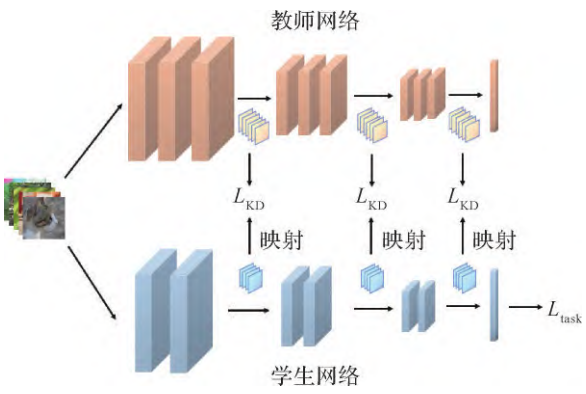

图 3　特征图上的知识蒸馏

Fig. 3　Knowledge distillation on feature maps

Zhi 等人（2017）将这种中间层特征图匹配的方式改为两阶段的形式，首先将教师网络的浅层部分和学生网络的深层部分连接起来，起到高维信息学习的效果，之后再完整地对学生网络进行蒸馏训练。类似地，Zhou 等人（2018）也尝试让教师网络和学生网络共享浅层网络结构。与前者不同，该方法让教师网络和学生网络同时进行训练，并且在网络输出部分使用蒸馏损失。

上述两种方法采用了直接让教师网络和学生网络共享特征图的方式，研究人员将这种共享结构扩展到了更大的范围。Fu 等人（2021）提出了交互式知识蒸馏，在这种模式下，学生网络和教师网络的结构会按照概率交替使用，而只针对对应的任务进行训练。这个思考的出发点是让学生网络在神经网络的各个位置上都接近教师网络。

直接拟合特征图是获取中间层信息最直接的方法，但同时也存在缺点。首先，这类方法对网络结构的要求较高。学生网络在拟合教师网络的特征时，需要尽量保证特征图大小相同，这样才能确保学生网络与教师网络推理过程的接近。其次，直接对特征图的拟合需要大量的计算，特征图的拟合结果与网络输出的拟合在数量级上有较大差别，难以确定损失函数与其他类型损失（任务损失和传统知识蒸馏损失）的比例，造成调参困难。

2.2.2　从特征图中提取知识

在直接拟合特征图的思路的基础上，研究者开始探索蕴含在中间层特征图中的知识，其中最直观的方法是使用数学方法从特征图中提取知识。Zagoruyko 和 Komodakis（2017）在提示层的基础上，使用了特征图的统计数据（通道维度上的最大值、平均值等）生成注意力图，以代替特征图的直接拟合。类似地，Yim 等人（2017）提出了另一种从特征图中提取信息的方法，将两层之间的特征图做内积，以获取特征图之间的“方向”信息，并将其作为知识指导学生网络的训练。

中间层获取的信息同样可以由学习来得到，一般的思路是将特征图的不同通道进行组合作为知识的表示形式，其损失函数的形式为

$$L_{\mathrm{KD}} = DF\left(\sum_i \alpha_i f_{\mathrm{S}}^i, \sum_j \beta_j f_{\mathrm{T}}^j\right) \tag{3}$$

式中，α 和 β 分别是学生网络和教师网络特征图中对应通道的权重，角标 i 和 j 是特征图对应的编号。

Zhang 等人（2020a）提出使用不同位置上的特征图进行对应任务的预测，如在用于分类任务的网络中加入额外的辅助分类器，使其能够捕捉到特征图上有关分类任务的信息，而这些辅助分类器的结果随后将用来进行知识蒸馏训练。

对于通道匹配的方法，Yue 等人（2020）根据教师网络和学生网络之间的相似性矩阵计算出二值的通道匹配矩阵，在通道数不同的教师—学生特征图对中建立多对一的映射，随后聚合教师网络的特征图以得到用于监督的特征图。Wang 等人（2020a）则采用学习的方法，加入了特征选择模块，分别从教师网络和学生网络中选取特征图进行匹配。

最新的研究工作开始对整个网络中不同位置的特征进行聚合，以达到充分利用教师网络中的知识的目的。该方法工作流程如图 4。Jang 等人（2019）引入了元学习的方法，使用元网络将网络中不同位置的特征图进行配对，计算出教师网络各层对应学生网络的权重，并加以整合计算损失。Ji 等人（2021a）采用自注意力机制（self-attention mechanism）的思路，根据教师网络特征图的位置对学生网络不同层次的特征图计算注意力权重，并根据权重对学生网络的特征进行聚合。从另一个方向，Chen 等人（2021）使用相同的思路将教师网络特征图进行特征聚合，作为知识对学生网络进行监督。对于上述方法中使用全部特征图进行聚合造成的冗余问题，Chen 等人（2021）提出只让学生网络学习教师网络中低于对应层次的信息，并采用渐进式的聚合方式获取教师网络中的知识。

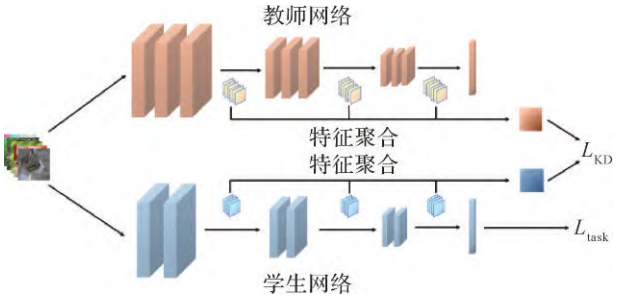

图 4　使用特征聚合的知识蒸馏

Fig. 4　Knowledge distillation with feature fusion

现有的从特征图中提取知识的方法虽然已经较为有效地解决了如何提取知识的问题，但是知识提取的位置仍然有待探索。当前的方法采用学习的思路寻找知识的位置，这种思路仍然存在着不可解释性和复杂性的问题，所以如何定位知识以及如何聚合知识仍然是今后的一个研究方向。

2.2.3　结构性知识

除了上述直接从特征图中提取知识的方法，研究者尝试通过分析网络的行为方式来获取知识，即结构性知识。常见的提取结构性信息的方式有样本间关系、类别间关系等，其损失函数的形式为

$$L_{\mathrm{KD}} = DF(M(f_{\mathrm{S}}^1, f_{\mathrm{S}}^2, \cdots), M(f_{\mathrm{T}}^1, f_{\mathrm{T}}^2, \cdots)) \tag{4}$$

式中，f 是特征图，其上角标表示样本的编号，而函数 $M(\cdot)$ 则用于计算网络对不同样本生成的特征图之间的距离。通过这些结构性知识，学生网络能够学习到教师网络对不同样本（或类别）不同的处理方式。这与传统的知识蒸馏是相似的，但是传统知识蒸馏只是隐式地让学生网络学习类间关系，而结构性知识则将这些知识显式地提取出来。图 5 以一次性输入 3 个样本的情况为例展示了使用结构性知识的知识蒸馏方法。

Park 等人（2019b）在网络输出层中提取样本间的知识，使用欧氏距离、角度距离等方式从多个样本组成的元组中提取样本之间的关系，并让学生网络对这些样本的输出之间的关系接近教师网络。Tung 和 Mori（2019）则站在特征图的角度，使用内积的方式计算一个批（batch）中所有样本的特征图的相似

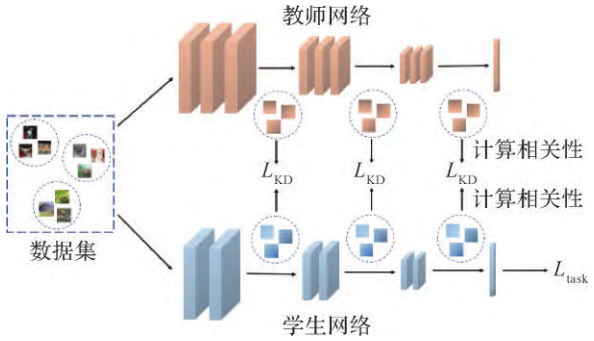

图 5　使用结构性知识的知识蒸馏

Fig. 5　Knowledge distillation with structural knowledge

度矩阵，并尝试让学生网络的特征图也得到类似的相似度关系。而 Liu 等人（2019a）则使用更加具体的方法分别为教师网络和学生网络建立了实例关系图（instance relationship graph，IRG），以节点代表特征图，以边代表特征图间的距离，让学生网络学习教师网络的实例关系图以及网络不同位置的实例关系图之间的关系，以学习到教师网络前向传播时的行为。

对比学习也是近期较为热门的研究热点，这个思路也被引入到知识蒸馏中来提取结构性知识。Tian 等人（2022）采用对比学习的思路，显式地建模出了在多种任务下的目标函数，即最大化教师网络和学生网络互信息的下界。Zhu 等人（2021）则关注于教师空间与教师—学生空间的一致性，分别用额外的结构提取二者在特征和梯度上的结构性知识，并最大化教师—学生之间的互信息。

在分类任务中，传统知识蒸馏可以将类别之间的信息突显出来供学生网络更好地学习，近期的一些工作通过显式地利用这里的类间信息达到增强教师网络指导的作用。Chen 等人（2020c）根据多个样本的特征图得到类内结构信息和类间结构信息，其中类内结构信息来自样本和类中心的距离，而类间结构则由映射函数学习得到，随后采用这些结构指导学生网络的行为。

当前使用结构性知识的方法已经能够做到显性地从教师网络向学生网络传递知识。但是由于涉及到多个样本间的操作和多个类别间的操作，这类方法同时也存在不易实现和训练的问题。

2.3　探索知识的学习方法

作为知识蒸馏的另一个重要组成部分，教师—学生架构也是研究重点之一。研究者从使用场景的角度出发，提出了一些不同于传统知识蒸馏教师—学生架构的方法，如在线蒸馏、多教师蒸馏等，以适应不同应用场景来提高学生网络的性能。本文将这些方法归类于对教师与学生之间学习的方法的探索，它们往往可以和各种知识的表示形式协同使用，以达到更好的效果。

2.3.1　在线蒸馏

在线蒸馏也称为深度相互学习（deep mutual learning，DML），最早由 Zhang 等人（2018）提出。在这个结构中不需要预训练的教师网络，而是使用多个学生网络同时进行学习，并使用知识蒸馏的方法互相学习对方学到的知识。在线蒸馏的损失函数可以表示为

$$L_{\mathrm{DML}}=\sum_{k}L_{\mathrm{t}}(\boldsymbol{l}_k,\boldsymbol{t})+\sum_{i<j}L_{\mathrm{KD}}(\boldsymbol{l}_i,\boldsymbol{l}_j) \tag{5}$$

式中，$\boldsymbol{l}$ 是各个子网络的输出，$\boldsymbol{t}$ 是任务目标，图 6 展示了在线蒸馏的模式，k 个学生网络的输出将两两形成损失。

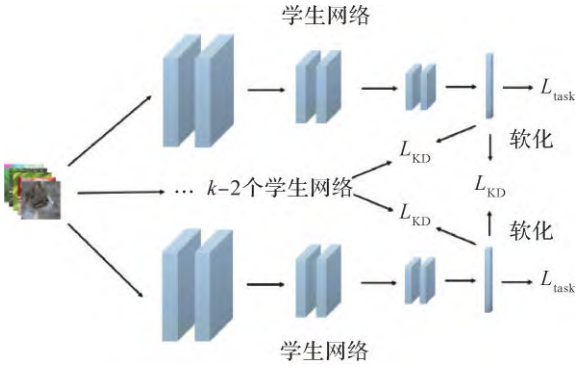

图 6　在线蒸馏

Fig. 6　Online distillation

与传统知识蒸馏类似，对在线蒸馏的改进主要关注于为多学生结构添加监督信息。Lan 等人（2018）使用学习的方法，将训练过程中学生网络的输出整合，形成一个在线的“教师网络”，使其能够在学习真实值的同时为所有的学生网络提供额外监督。类似地，Chen 等人（2020a）向在线蒸馏的结构中添加了合成的教师网络提供额外监督。与前者不同，该方法设置了两层蒸馏，第 1 层蒸馏使用非对称的自注意力机制为每个学生网络生成一个监督信息，而第 2 层蒸馏将第 1 层的监督信息整合为单独的一个学生网络提供监督，从而提供了多样化的监督信息。循着同样的思路，Kim 等人（2021）在此基础上添加了特征聚合模块，将学生网络的特征进行聚

合以提供“教师网络”的监督。

在线蒸馏实现了不依赖预训练的教师网络和提升网络性能两个目标。但是新的研究发现这个框架中仍然需要一个合成的教师网络以进一步提升网络性能，这样就会因同时训练多种网络而产生难以训练的问题。

2.3.2 自蒸馏

自蒸馏（self-distillation，SD）指在网络训练过程中利用自身已经学习到的信息优化自身的学习过程，可以看做是一个用于提升网络性能的方法。与在线蒸馏类似，自蒸馏可以在没有教师网络的情况下使用。相对于传统的知识蒸馏，自蒸馏省去了对教师网络的选择和训练的过程，但是同样可以达到提升网络性能的目的。自蒸馏的损失函数可以表示为

$$L_{SD}=L_t(\boldsymbol{l},\boldsymbol{t})+\sum_{i=1}^{N-1}DF(\boldsymbol{f}_i,R(\boldsymbol{f}_{i+1})) \tag{6}$$

式中，$\boldsymbol{f}_i$表示网络中第i个部分的特征图，$R(\cdot)$是由深层次特征图重建出浅层次特征图的函数，自蒸馏过程如图7所示。

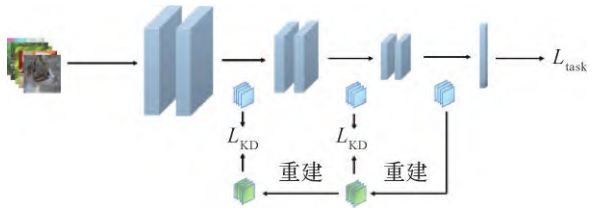

图7 自蒸馏

Fig. 7 Self-distillation

自蒸馏最早出现在Zhang等人（2019）的工作中，在神经网络的每个阶段都添加了输出层，并且使用多个来自高层特征的监督信息，包括特征图的提示监督、输出层的知识蒸馏监督以及真实标签的监督。在此基础上，Liu等人（2020a）加入了元学习的思路，使用元网络根据深层特征生成浅层网络对应的监督信息，从而实现自顶向下的自监督学习。Ji等人（2021b）则使用自身网络的特征自下而上地生成了对各个阶段的监督，并同时加入自上而下的特征以增强这些监督信号，从而丰富自监督的知识。

自蒸馏学习中的监督不一定来自自身的特征图，也可以来自其他结构化的信息。Ge等人（2021）将样本间相似性作为监督信息，计算选定样本和批（batch）中其他所有样本的相似度，并根据结果对输出进行加权得到最终的监督信息，随后在对选定的样本的前向传播中进行监督。

类似在线蒸馏，自蒸馏模式同样达到了脱离预训练模型的依赖的目的，但是同样存在一个明显的弊端，即无法确保自身保留的知识是否准确，导致这类方法训练存在随机性较强的问题。

2.3.3 多教师蒸馏

为了弥补单个教师网络能够提供的监督有限的问题，一些工作采用多个教师网络增加更多的监督信息，即多教师蒸馏（multi-teacher distillation，MTD）。多教师蒸馏的优势在于，除了能够提供传统知识蒸馏中的软化标签之外，还能提供更加多样化的信息，但存在获取教师网络难度高的缺陷。多教师蒸馏的损失函数可以表示为

$$L_{MTD}=L_t(\boldsymbol{l}_S,\boldsymbol{t})+L_{KD}(\boldsymbol{l}_S,E(\boldsymbol{l}_{T1},\boldsymbol{l}_{T2},\cdots)) \tag{7}$$

式中，$E(\cdot)$表示对教师网络的输出整合函数。多教师蒸馏过程如图8所示。

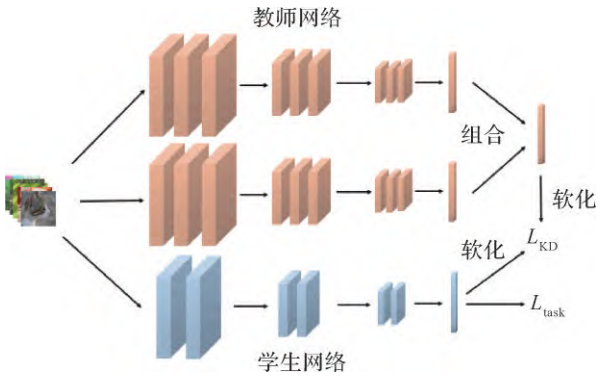

图8 多教师蒸馏

Fig. 8 Multi-teacher distillation

早期多教师蒸馏的工作中，You等人（2017）采用平均所有教师网络软标签以及最大化样本间相似度的损失，将教师网络的信息进行整合，并用做知识蒸馏的监督。之后的研究关注点主要在如何选择教师网络上，Park等人（2019a）加入了一系列的非线性变换，对学生网络特征和教师网络特征进行匹配并学习对应的特征。Liu等人（2020d）采用注意力机制，计算教师网络特征与学生网络特征的内积以得到匹配权重，并同时加入提示损失和基于结构性知识的角度损失。

Asif等人（2020）将研究目标指向多教师多学生的蒸馏模式，通过分别对教师网络和学生网络的特征图进行整合，让学生网络分别学习不同教师网络

的多样化知识，并最终加以整合。

多教师蒸馏的模式可以为学生网络提供更加丰富和多样的监督信息，但是同样也增加了知识蒸馏方法的使用代价。如何选择合适的教师网络，以及如何训练教师网络供多样化的知识，这些问题都有待进一步研究。

2.3.4　渐进式蒸馏

渐进式蒸馏（progressive knowledge distillation，PKD）的思路是使用教师网络对学生网络的训练过程进行监督，而非直接提供一个预训练模型的输出，这有助于尚未获得足够知识的学生网络了解到教师网络学习知识的过程。渐进式蒸馏的损失函数为

$$L_{\mathrm{PKD}} = L_{\mathrm{t}}(\boldsymbol{l}_{\mathrm{S}},\boldsymbol{t}) + L_{\mathrm{t}}(\boldsymbol{l}_{\mathrm{T}},\boldsymbol{t}) + L_{\mathrm{KD}}(\boldsymbol{l}_{\mathrm{S}},\boldsymbol{l}_{\mathrm{T}}) \tag{8}$$

上述形式描述的是教师和学生网络同时训练的情形。

Yang 等人（2019）提出使用教师网络训练过程中的一系列“快照”（snapshot），分阶段对学生网络进行监督。而 Zhao 等人（2019）更加细化了这种渐进式的监督，使用一个预训练的教师网络和一个跟随学生网络训练的教师网络，二者分别在中间层和输出层给予学生网络指导。Shi 等人（2021）则舍弃了预训练模型，使得这种训练方法的应用范围更广。具体来说，将教师网络和学生网络进行交替训练，同时使用额外的损失限制了教师网络和学生网络性能的差距，以使指导更为有效。

渐进式蒸馏的方法细化了知识蒸馏的训练过程，给出了学生网络阶段性的学习目标。这种思路降低了学生网络的学习难度，但是在多阶段的神经网络训练中（如训练期间调整学习率、交替优化多个网络部分等），其不确定性会增大，从而导致训练效果的不稳定。

2.3.5　基于生成对抗网络的知识蒸馏

知识蒸馏的“教师—学生”架构和生成对抗网络（generative adversarial networks，GAN）的“生成器—判别器”架构具有很高的相似性，教师网络的输出可以看做“真实值”，而学生网络的输出可以看做“生成值”。根据这种博弈的思想，学生网络将尝试骗过判别器，而判别器尝试将教师网络的输出和学生网络的输出分开，以促使学生网络获得更加接近教师网络的能力。基于 GAN 的知识蒸馏可以表达为

$$L_{\mathrm{GAN-KD}} = L_{\mathrm{t}}(\boldsymbol{l}_{\mathrm{S}},\boldsymbol{t}) + L_{\mathrm{GAN}}(\boldsymbol{l}_{\mathrm{S}},\boldsymbol{l}_{\mathrm{T}},D) \tag{9}$$

式中，D 表示判别器，L_{GAN} 是训练判别器使用的损失，具体形式为

$$L_{\mathrm{GAN}}(\boldsymbol{l}_{\mathrm{S}},\boldsymbol{l}_{\mathrm{T}},D) = \log D(\boldsymbol{l}_{\mathrm{T}}) + \log D(\boldsymbol{1} - \boldsymbol{l}_{\mathrm{S}}) \tag{10}$$

基于 GAN 的知识蒸馏过程如图 9 所示。Xu 等人（2018）在知识蒸馏中使用了条件生成对抗网络（conditional generative adversarial networks，C-GAN）的思路，让判别器在分辨教师网络输出和学生网络输出的同时也对类别进行预测，从而首次做到了 GAN 在提升知识蒸馏性能上的成功应用。而随后 Chen 等人（2020b）则使用 GAN 的判别器对教师网络和学生网络的特征图进行判别。Chung 等人（2020）将 GAN 的思路带到了在线蒸馏，在相邻的两个学生网络之间添加一个判别器，促使各个子网络保留丰富的多样性。黄仲浩等人（2022）尝试在 GAN 的结构中细化教师网络对学生网络的监督，采用逐层贪婪的监督策略，最大化利用教师网络的监督信息，并在判别器的额外监督下达到了较好的效果。

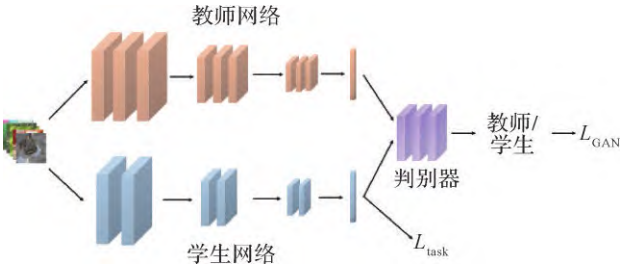

图 9　基于 GAN 的知识蒸馏

Fig. 9　GAN-based knowledge distillation

3　面向网络压缩的知识蒸馏

知识蒸馏作为神经网络压缩方法，在网络压缩效率方面的研究层出不穷。对于在知识蒸馏中考虑网络规模的情形，常见的两个思路是对网络进行动态调整，即网络架构搜索（NAS）以及尝试将知识蒸馏推广到更小规模的网络上。这两个思路分别从动态调整网络结构和网络结构的选择选择两个方面考虑神经网络的规模。

3.1　动态调整网络结构

将知识蒸馏与 NAS 结合是一个十分直观的思路，可以使网络在蒸馏训练的过程中调整结构，以达到性能和网络规模的平衡。

Mitsuno 等人（2021）提出一种类似 NAS 思路的简单的知识蒸馏方法，由一个规模很小的学生网络

开始，将网络中的层分为若干组，每一轮迭代用穷举的方法对每一组中的所有层添加一定数量的通道数，并选出其中性能最好的网络，直到不再产生比上一轮迭代性能更好的网络为止。Liu等人(2020b)将NAS完整地引入知识蒸馏中，使用强化学习的方式，根据知识蒸馏训练和网络在移动端上的使用延迟给出反馈，并在一个较大的搜索空间中进行搜索。

同时，知识蒸馏的思路也可以辅助神经网络剪枝方法。Li等人(2020b)首先采用剪枝的方法得到学生网络，再在学生网络中插入卷积层，以对齐教师网络与学生网络的特征，并使用无标签数据进行中间层的知识蒸馏训练，达到恢复网络性能的目的。

3.2 小规模网络蒸馏训练

在神经网络压缩的应用场景下，研究者希望“教师—学生”架构中的学生网络有较小的规模。但是，当教师网络和学生网络的规模差异过大时，蒸馏的效果会受到明显影响(Cho和Hariharan，2019；Mirzadeh等，2019)，这是因为学生网络的容量不足以学习到教师网络的知识。为了将知识蒸馏训练推广到更小的网络架构上，Cho和Hariharan(2019)提出用于知识蒸馏的提前停止(early stopping)策略，在训练早期使用知识蒸馏训练学生网络，而在进行到一定程度后停止教师网络的监督，使用真实标签进行后续训练。

另外，Mirzadeh等人(2019)采用一种新的策略，即在差别较大的教师网络和学生网络中间插入“助教”网络，让规模介于二者之间的助教网络学习到教师网络的知识后再传递给学生网络，并提出了蒸馏路径上助教网络越多蒸馏效果越好的假设。在此基础上，Son等人(2021)尝试解决助教网络较多的情况下错误传播放大的问题，提出对助教网络或学生网络训练时，使用所有规模大于它的网络进行蒸馏训练，并加入用于助教网络之间的随机失活策略(dropout)来解决计算量过大的问题。类似地，刘昊和张晓滨(2021)提出了基于关系型蒸馏的分步神经网络压缩方法，通过在每一层知识蒸馏中加入额外的监督信息，达到提升小规模网络性能的效果。

4 知识蒸馏方法性能对比

知识蒸馏方法尝试将教师网络从数据集中学习到的知识以各种形式传递给学生网络，以达到将学生网络的性能提升至接近教师网络的目的。本文对不同类别的知识蒸馏方法中具有代表性的方法进行实验对比，以对不同知识蒸馏方法提性能提升的效果进行直观可视化。

知识蒸馏作为辅助训练型方法，训练所用数据集和训练方法均与对应任务相同。所有实验均在分类任务常用的CIFAR10(Canadian Institute for Advanced Research)数据集上进行，该数据集的图像均为32 × 32像素，共有50 000个训练集数据和10 000个测试集数据。为了公平对比，统一采用ResNet56作为教师网络(如果使用了教师网络)，ResNet20作为学生网络，并采用同样的超参数(包括学习率、学习率衰减和温度参数等)。实验采用3次实验的分类准确率作为最终的性能评价标准。实验结果如表1所示。

从实验结果可得出以下结论：1)整体而言，当前所有的知识蒸馏方法都能够使学生网络达到比直接训练更好的性能。但是由于网络规模的差异过大，这些方法都未能达到教师网络的性能水平；2)参与实验对比的方法中，在线蒸馏系列方法能够达到最好效果，且普遍高于传统的使用网络输出进行一对一蒸馏的方法；3)使用结构化知识的效果优于直接从网络中提取知识的方法，这是由于使用结构化知识的方法更倾向于学习教师网络的行为特征。

5 知识蒸馏方法的应用

知识蒸馏的应用也可按上述方法分类，即使用知识蒸馏的思路提升网络在该任务上的性能，以及使用知识蒸馏方法得到轻量化的网络。本文列举了在多个任务下使用知识蒸馏进行性能提升和网络压缩的方法。

5.1 目标检测

目标检测任务是对轻量级网络需求较大的任务之一，在目标检测中可以直接使用知识蒸馏的思路来获取轻量级网络，如Chen等人(2017)使用一个大规模的目标检测网络作为教师网络，从主干网络、区域提议网络(region proposal network，RPN)和区域分类网络(region classification network，RCN)3个部分进行蒸馏训练。Wei等人(2018)则在前者的基础上对教师网络和学生网络的特征图进行量化，之后再进行蒸馏训练，这样可以有效减小学习的难度，并使

表1 知识蒸馏方法在CIFAR10数据集上的性能对比

Table 1 Performance comparison of different KD methods on CIFAR10 dataset

类别	方法	分类准确率/%	备注
直接训练	ResNet56(He等,2016)	94.01	
	ResNet20(He等,2016)	92.36	
传统方式	KD(Hinton等,2015)	92.43	
	CRD(Tian等,2022)	92.69	
	TAKD (Mirzadeh等,2019)	92.83	ResNet32作为助教网络
直接拟合特征图	FitNet(Romero等,2015)	92.86	
特征图提取知识	AT (Zagoruyko和Komodakis,2017)	92.89	
	AFD (Wang等,2020a)	92.90	
	FSP(Yim等,2017)	92.48	
结构化知识	RKD(Park等,2019b)	92.89	
	IRG(Liu等,2019a)	92.95	
	SPKD(Tung和Mori,2019)	93.14	
在线蒸馏	DML(Zhang等,2018)	92.93	3个学生网络
	ONE(Lan等,2018)	93.08	3个学生网络
	OKDDip(Chen等,2020a)	93.24	3个学生网络
多教师蒸馏	ALTKD(Liu等,2020d)	92.53	3个教师网络

量化后的学生网络得到较好的性能。

Dai等人(2021)采用更加丰富的蒸馏监督,使用筛选模块选择出最有价值的检测框,之后计算学生网络和教师网络在对应检测框中特征图间的关系,与应用于检测结果中的知识蒸馏共同组成完整的训练损失。类似地,在Yao等人(2021)的工作中也采用检测框特征学习的思路,提取检测框在特征金字塔中对应的特征图,并使用自注意力机制对金字塔中各层特征图进行整合,随后再学习教师网络对应的特征图。楚玉春等人(2022)则同时使用检测框的知识和目标分类中的知识,并采用多层注意力机制的方法提升知识传输的效率,从而得到更好的训练效果。

还有一些方法使用类似知识迁移的思路,侧重于某项特定知识的提取和学习。Li等人(2017)尝试使用在ImageNet数据集上预训练的分类模型作为教师网络,并用一个轻量级的网络从输出和中间层学习教师网络学到的知识,以在基本不损失性能的情况下大量减小使用代价。Valverde等人(2021)使用多模态的信息,如彩色图像、热力图像和深度图等,用对应于3种图像的网络的检测结果监督音频信息,并最终使用音频信息得到多种图像上的检测结果。

5.2 语义分割

语义分割相对于目标检测更倾向于细粒度的知识(如像素级等),同样适合作为知识蒸馏的应用场景。Liu等人(2019b)使用两个网络特征图之间的距离作为逐对蒸馏(pair-wise distillation)损失,计算二者最终输出的像素分数的相似性,形成逐像素蒸馏(pixel-wise distillation)损失,并根据GAN的思路加入判别器分辨教师网络和学生网络的输出。Shu等人(2020)则将直接对特征图进行的蒸馏换成了逐通道蒸馏(channel-wise distillation),并将这种思路同时用在了分割结果的蒸馏上,以使得学生网络的每个通道获取更加具有特异性的知识。

对于域适应场景分割的应用,Kothandaraman等人(2021)分别在教师网络和学生网络上建立源域到目标域的知识蒸馏策略,结合主干网络中特征图的特征和分割结果的特征以完成知识蒸馏训练。

在图像的协同分割任务中,耿增民等人(2020)将知识蒸馏应用在孪生神经网络中以达到降低网络复杂度的目的。他们尝试使用二值化注意力机制的

方式将教师网络中的知识提取出来，并依此重构得到学生网络结构，随后使用知识蒸馏训练，得到接近教师网络性能的学生网络模型。

5.3 人脸识别

人脸识别任务相对于分类等任务，需要更多地考虑人脸姿态和遮挡等问题，而知识蒸馏能够提供这方面的支持以提高人脸识别的效果。Huang等人(2020)尝试使用知识蒸馏的思路对人脸特征的分布关系进行蒸馏，将简单样本正负对的相似度关系视为教师分布，将困难样本正负对的相似度关系作为学生分布，将二者进行拟合以达到提升网络对困难样本进行识别的效果。

对于网络轻量化，Shi等人(2020)使用较为直接的蒸馏方式，将单个或多个教师网络分类器部分的权重直接使用在学生网络中，并采用人脸角度损失(arcface loss)进行蒸馏训练。Wang等人(2020c)定义神经网络权重特异性(weight exclusivity)以描述权重的多样性，并提出使用特异性作为正则项加入到知识蒸馏的训练中，以保证在人脸识别任务中学生网络能够保留更丰富的特征。

5.4 图像生成

在图像生成类任务中，生成对抗网络是最常用的方法，而GAN中使用的生成器往往会有较大的使用代价，所以对GAN的压缩也是一个常见的应用，但是GAN生成器独特的结构特性使其不能直接使用传统的知识蒸馏进行压缩。Chen等人(2020b)使学生网络生成器生成的图像在像素值上和语义上学习教师网络生成的图像，并将教师网络生成的图像作为真实值在学生网络的判别器中进行训练。

Ren等人(2021)尝试利用多粒度的信息对GAN进行压缩，采用在线蒸馏的形式，根据学生网络的结构定制出两个不同的教师网络以获取不同层次的信息，并同时进行训练，在这个结构中，学生网络不需要判别器进行训练，而是采用图像质量损失、图像风格损失和感知损失直接进行知识蒸馏训练。

为了更好地达到规模与性能的平衡，Li等人(2020a)尝试使用NAS方法对条件GAN进行压缩，采用通道剪枝的方法得到学生网络结构，并使用搜索方法确定通道数量的最佳配置。在训练时采用“一劳永逸”(once-for-all)的NAS方法将训练和搜索进行解耦，并在训练中加入了基于中间层信息的知识蒸馏。Jin等人(2021)尝试在GAN和自编码器的设定下，使用教师网络同时完成知识蒸馏的任务和网络架构搜索的任务，参照GoogLeNet的形式将教师网络的基本模块设计成多路特征的形式，并使用剪枝的方式进行架构搜索得到学生网络。

5.5 自然语言处理

除了研究广泛的计算机视觉(computer vision，CV)任务，知识蒸馏也用于自然语言处理(natural language processing，NLP)任务中，以获得更加轻量级的网络。

BERT(bidirectional encoder representation from Transformers)是NLP任务中的常用结构，由多个Transformer结构堆叠而来，在具有较好性能的同时也存在规模较大的问题。Sun等人(2019)在对BERT的蒸馏中使用了对输出的监督和对中间层堆叠的Transformer的监督，同时为了避免中间层计算量过大，提出让学生网络在此过程中只对特殊字段进行学习。随后Jiao等人(2020)开发了tinyBERT结构，在输出和中间层监督的基础上添加了Transformer内部的知识蒸馏，包括多头注意力模块以及隐藏状态上的监督，同时提出了两段式的学习框架，即首先在大规模语料库下进行知识蒸馏训练，之后再针对特定的任务进行知识蒸馏训练得到最终的模型。

Fu等人(2021)引入了对比学习的方式对BERT进行知识蒸馏训练，尝试让学生网络的中间层特征接近对应样本在教师网络中的特征，而远离其他样本在教师网络中的特征，并使用更加适合度量语义信息的基于角度的距离来进行距离的度量。

6 总结与展望

近年知识蒸馏方法取得了显著发展，在神经网络性能提升和神经网络压缩方面都形成了体系化的发展方向，在各大任务上得到了广泛应用。但是知识蒸馏方法仍然存在需要解决的问题，下面从3个方面对知识蒸馏的现状及问题进行总结，并对其未来发展进行展望。

1)知识蒸馏的研究主要集中于网络性能提升的目标，在这个目标之下，研究者探究了知识的表示形式以及知识的学习方式。这类方法针对的主要问题是知识的来源，即确定何种知识是有效的，这也是今后知识蒸馏方法研究的一个主要问题。当前对知识

的表示形式的研究主要局限于同种任务网络的知识以及样本之间的知识，所以探索其他形式的网络（如生成式网络等）和其他任务（如语义分割等）中可用的信息，并引入到其他任务中是一个可行的研究思路。

2）在以神经网络压缩为目标的研究中，最常采用的是神经架构搜索的思路。但是现有的神经架构搜索方法往往存在使用代价巨大的问题，再加上知识蒸馏需要额外的代价训练教师网络，减小 NAS 方法的搜索空间使之适合知识蒸馏任务成为一个可行的研究思路。同时，根据教师—学生架构的相关研究，随着网络结构的变化，对教师网络进行调整也是一个可行的研究方向。

3）除了研究方法上的创新，包括知识蒸馏在内的神经网络压缩方法还需要一个具有普遍意义的衡量标准。在以性能提升为目标的方法中，研究者采用的评价标准往往是学生网络的性能变化，而在以网络压缩为目的的方法中，则常常是将压缩率和性能变化直接以某种规律排列起来，而使用这种方法进行比较无法直观地对比各个方法的效果，所以制订一个较为统一的评价标准也是未来的研究方向之一。

7 结 语

随着深度学习方法效果的不断提升，应用代价高的问题也越来越显著，而使用神经网络压缩方法则是解决这一问题的有效途径。知识蒸馏作为神经网络压缩中的重要方法，自提出以来就受到广泛关注。本文从知识蒸馏方法的目标的角度（网络性能提升及网络压缩）对知识蒸馏方法进行了分类和归纳，并列举了知识蒸馏在多种任务中的应用场景。

根据本文的归纳与分析，在近年来的知识蒸馏方法中，面向网络性能的知识蒸馏是主流研究方向，包括对中间层知识、结构性知识的探索，以及对无教师、多教师等情形下的研究。实验表明，当前的知识蒸馏方法已经能够为神经网络训练提供较好的性能提升效果，使其能够代替大规模网络应用于图像分类等任务中。在此基础上，结合多个任务中的知识，以及设计能够适用于知识蒸馏的神经架构搜索空间，以获得更好的性能与压缩率的平衡，将是未来知识蒸馏可行的研究方向。

参考文献（References）

Asif U, Tang J B and Harrer S. 2020. Ensemble knowledge distillation for learning improved and efficient networks//Proceedings of the 24th European Conference on Artificial Intelligence. Santiago de Compostela, Spain: IOS Press: 953-960 [DOI: 10.3233/FAIA200188]

Chen D, Mei J P, Wang C, Feng Y and Chen C. 2020a. Online knowledge distillation with diverse peers. Proceedings of the AAAI Conference on Artificial Intelligence, 34(4): 3430-3437 [DOI: 10.1609/aaai.v34i04.5746]

Chen G B, Choi W, Yu X, Han T and Chandraker M. 2017. Learning efficient object detection models with knowledge distillation//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 742-751

Chen H T, Wang Y H, Shu H, Wen C Y, Xu C J, Shi B X, Xu C and Xu C. 2020b. Distilling portable generative adversarial networks for image translation. Proceedings of the AAAI Conference on Artificial Intelligence, 34(4): 3585-3592 [DOI: 10.1609/AAAI.V34I04.5765]

Chen P G, Liu S, Zhao H S and Jia J Y. 2021. Distilling knowledge via knowledge review//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 5006-5015 [DOI: 10.1109/CVPR46437.2021.00497]

Chen Z L, Zheng X X, Shen H L, Zeng Z Y, Zhou Y K and Zhao R C. 2020c. Improving knowledge distillation via category structure//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 205-219 [DOI: 10.1007/978-3-030-58604-1_13]

Cho J H and Hariharan B. 2019. On the efficacy of knowledge distillation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 4793-4801 [DOI: 10.1109/ICCV.2019.00489]

Chu Y C, Gong H, Wang X F and Liu P S. 2022. Study on knowledge distillation of target detection algorithm based on YOLOv4. Computer Science, 49(6A): 337-344 (楚玉春, 龚航, 王学芳, 刘培顺. 2022. 基于YOLOv4的目标检测知识蒸馏算法研究. 计算机科学, 49(6A): 337-344) [DOI: 10.11896/jsjkx.210600204]

Chung I, Park S U, Kim J and Kwak N. 2020. Feature-map-level online adversarial knowledge distillation [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/2002.01775.pdf

Dai X, Jiang Z R, Wu Z, Bao Y P, Wang Z C, Liu S and Zhou E J. 2021. General instance distillation for object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 7838-7847 [DOI:

10.1109/CVPR46437.2021.00775]

Fu H, Zhou S J, Yang Q H, Tang J J, Liu G Q, Liu K K and Li X L. 2021. LRC-BERT: latent-representation contrastive knowledge distillation for natural language understanding. Proceedings of the AAAI Conference on Artificial Intelligence, 35(14): 12830-12838[DOI: 10.1609/aaai.v35i14.17518]

Gao M Y, Shen Y J, Li Q Q, Yan J J, Wan L, Lin D H, Loy C C and Tang X O. 2019. An embarrassingly simple approach for knowledge distillation[EB/OL].[2022-03-13]. https://arxiv.org/pdf/1812.01819v2.pdf

Ge Y X, Zhang X, Choi C L, Cheung K C, Zhao P P, Zhu F, Wang X G, Zhao R and Li H S. 2021. Self-distillation with batch knowledge ensembling improves ImageNet classification[EB/OL].[2022-03-13]. https://arxiv.org/pdf/2104.13298.pdf

Geng Z M, Yu M Q, Liu X B and Lyu C. 2020. Combining attention mechanism and knowledge distillation for Siamese network compression. Journal of Image and Graphics, 25(12): 2563-2577(耿增民, 余梦巧, 刘峡壁, 吕超. 2020. 融合注意力机制与知识蒸馏的孪生网络压缩. 中国图象图形学报, 25(12): 2563-2577)[DOI: 10.11834/jig.200051]

Gou J P, Yu B S, Maybank S J and Tao D C. 2021. Knowledge distillation: a survey. International Journal of Computer Vision, 129(6): 1789-1819[DOI: 10.1007/s11263-021-01453-z]

He K, Zhang X, Ren S and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 770-778[DOI: 10.1109/CVPR.2016.90]

Hinton G, Vinyals O and Dean J. 2015. Distilling the knowledge in a neural network[EB/OL].[2022-03-13]. https://arxiv.org/pdf/1503.02531.pdf

Huang Y G, Shen P C, Tai Y, Li S X, Liu X M, Li J L, Huang F Y and Ji R R. 2020. Improving face recognition from hard samples via distribution distillation loss//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 138-154[DOI: 10.1007/978-3-030-58577-8_9]

Huang Z H, Yang X Y, Yu J, Guo L and Li X. 2022. Mutual learning knowledge distillation based on multi-stage Multi-generative Adversarial Network. Computer Science, 1-12(黄仲浩,杨兴耀,于炯,郭亮,李想.基于多阶段多生成对抗网络的互学习知识蒸馏方法.计算机科学:1-12)[DOI: 10.11896/jsjkx.210800250]

Jang Y, Lee H, Hwang S J and Shin J. 2019. Learning what and where to transfer//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 3030-3039

Ji M, Heo B and Park S. 2021a. Show, attend and distill: knowledge distillation via attention-based feature matching. Proceedings of the AAAI Conference on Artificial Intelligence, 35(9): 7945-7952[DOI: 10.1609/aaai.v35i9.16969]

Ji M, Shin S, Hwang S, Park G and Moon I C. 2021b. Refine myself by teaching myself: feature refinement via self-knowledge distillation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 10659-10668[DOI: 10.1109/CVPR46437.2021.01052]

Jiao X Q, Yin Y C, Shang L F, Jiang X, Chen X, Li L L, Wang F and Liu Q. 2020. TinyBERT: distilling BERT for natural language understanding//Findings of the Association for Computational Linguistics: EMNLP 2020. Virtual: ACL: 4163-4174[DOI: 10.18653/v1/2020.findings-emnlp.372]

Jin Q, Ren J, Woodford O J, Wang J Z, Yuan G, Wang Y Z and Tulyakov S. 2021. Teachers do more than teach: compressing image-to-image models//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13595-13606[DOI: 10.1109/CVPR46437.2021.01339]

Kim J, Hyun M, Chung I and Kwak N. 2021. Feature fusion for online mutual knowledge distillation//Proceedings of the 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE: 4619-4625[DOI: 10.1109/ICPR48806.2021.9412615]

Kothandaraman D, Nambiar A and Mittal A. 2021. Domain adaptive knowledge distillation for driving scene semantic segmentation//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW). Waikola, USA: IEEE: 134-143[DOI: 10.1109/WACVW52041.2021.00019]

Lan X, Zhu X T and Gong S G. 2018. Knowledge distillation by on-the-fly native ensemble//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.: 7528-7538

LeCun Y, Bengio Y and Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444[DOI: 10.1038/nature14539]

Li M Y, Lin J, Ding Y Y, Liu Z J, Zhu J Y and Han S. 2020a. GAN compression: efficient architectures for interactive conditional GANs//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 5283-5293[DOI: 10.1109/CVPR42600.2020.00533]

Li Q Q, Jin S Y and Yan J J. 2017. Mimicking very efficient network for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 7341-7349[DOI: 10.1109/CVPR.2017.776]

Li T H, Li J G, Liu Z and Zhang C S. 2020b. Few sample knowledge distillation for efficient network compression//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: 14627-14635[DOI: 10.1109/cvpr42600.2020.01465]

Li X J, Wu J L, Fang H Y, Liao Y, Wang F and Qian C. 2020c. Local correlation consistency for knowledge distillation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 18-33[DOI: 10.1007/978-3-030-58610-2_2]

Liu B L, Rao Y M, Lu J W, Zhou J and Hsieh C J. 2020a. MetaDistiller: network self-boosting via meta-learned top-down distillation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 694-709 [DOI: 10.1007/978-3-030-58568-6_41]

Liu H and Zhang X B. 2021. Compression method for stepwise neural network based on relational distillation. Computer Systems and Applications, 30(12): 248-254 (刘昊, 张晓滨. 2021. 基于关系型蒸馏的分步神经网络压缩方法. 计算机系统应用, 30(12): 248-254) [DOI: 10.15888/j.cnki.csa.008202]

Liu Y, Jia X H, Tan M X, Vemulapalli R, Zhu Y K, Green B and Wang X G. 2020b. Search to distill: pearls are everywhere but not the eyes//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 7536-7545 [DOI: 10.1109/CVPR42600.2020.00756]

Liu Y, Zhang W and Wang J. 2020c. Learning from a lightweight teacher for efficient knowledge distillation [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/2005.09163v1.pdf

Liu Y A, Zhang W and Wang J. 2020d. Adaptive multi-teacher multi-level knowledge distillation. Neurocomputing, 415: 106-113 [DOI: 10.1016/j.neucom.2020.07.048]

Liu Y F, Cao J J, Li B, Yuan C F, Hu W M, Li Y X and Duan Y Q. 2019a. Knowledge distillation via instance relationship graph//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Lone Beach, USA: IEEE: 7089-7097 [DOI: 10.1109/CVPR.2019.00726]

Liu Y F, Chen K, Liu C, Qin Z C, Luo Z B and Wang J D. 2019b. Structured knowledge distillation for semantic segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 2599-2608 [DOI: 10.1109/CVPR.2019.00271]

Meng Z, Li J Y, Zhao Y and Gong Y F. 2019. Conditional teacher-student learning//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK: IEEE: 6445-6449 [DOI: 10.1109/ICASSP.2019.8683438]

Mirzadeh S I, Farajtabar M, Li A and Ghasemzadeh H. 2019. Improved knowledge distillation via teacher assistant: bridging the gap between student and teacher [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/1902.03393v1.pdf

Mitsuno K, Nomura Y and Kurita T. 2021. Channel planting for deep neural networks using knowledge distillation//Proceedings of the 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE: 7573-7579 [DOI: 10.1109/ICPR48806.2021.9412760]

Oki H, Abe M, Miyao J and Kurita T. 2020. Triplet loss for knowledge distillation//Proceedings of 2020 International Joint Conference on Neural Networks. Glasgow, UK: IEEE: 1-7 [DOI: 10.1109/IJCNN48605.2020.9207148]

Park S U and Kwak N. 2019a. FEED: feature-level ensemble for knowledge distillation [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/1909.10754.pdf

Park W, Kim D, Lu Y and Cho M. 2019b. Relational knowledge distillation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 3962-3971 [DOI: 10.1109/CVPR.2019.00409]

Ren Y X, Wu J, Xiao X F and Yang J C. 2021. Online multi-granularity distillation for GAN compression [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/2108.06908.pdf

Romero A, Ballas N, Kahou S E, Chassang A, Gatta C and Bengio Y. 2015. FitNets: hints for thin deep nets [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/1412.6550.pdf

Shi W D, Ren G H, Chen Y P and Yan S C. 2020. ProxylessKD: direct knowledge distillation with inherited classifier for face recognition [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/2011.00265.pdf

Shi W X, Song Y X, Zhou H, Li B H and Li L. 2021. Follow your path: a progressive method for knowledge distillation//Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Bilbao, Spain: Springer: 596-611 [DOI: 10.1007/978-3-030-86523-8_36]

Shu C Y, Liu Y F, Gao J F, Xu L and Shen C H. 2020. Channel-wise distillation for semantic segmentation [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/2011.13256v1.pdf

Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/1409.1556v6.pdf

Son W, Na J, Choi J and Hwang W. 2021. Densely guided knowledge distillation using multiple teacher assistants [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/2009.08825v1.pdf

Sun S Q, Cheng Y, Gan Z and Liu J J. 2019. Patient knowledge distillation for BERT model compression//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: ACL: 4323-4332 [DOI: 10.18653/v1/D19-1441]

Szegedy C, Liu W and Jia Y. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]

Tian Y L, Krishnan D and Isola P. 2022. Contrastive representation distillation [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/1910.10699.pdf

Tung F and Mori G. 2019. Similarity-preserving knowledge distillation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 1365-1374 [DOI: 10.1109/ICCV.2019.00145]

Valverde F R, Hurtado J V and Valada A. 2021. There is more than

meets the eye: self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 11607-11616 [DOI: 10.1109/CVPR46437.2021.01144]

Wang K F, Gao X T, Zhao Y R, Li X J, Dou D J and Xu C Z. 2020a. Pay attention to features, transfer learn faster CNNs//Proceedings of the 8th International Conference on Learning Representations (ICLR). Addis Ababa, Ethiopia: Open Review:1-14

Wang X B, Fu T Y, Liao S C, Wang S, Lei Z and Mei T. 2020c. Exclusivity-consistency regularized knowledge distillation for face recognition//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 325-342 [DOI: 10.1007/978-3-030-58586-0_20]

Wei Y, Pan X Y, Qin H W, Ouyang W L and Yan J J. 2018. Quantization mimic: towards very tiny CNN for object detection//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 274-290 [DOI: DOI: 10.1007/978-3-030-01237-3_17]

Wen T C, Lai S Q and Qian X M. 2021. Preparing lessons: improve knowledge distillation with better supervision. Neurocomputing, 454: 25-33 [DOI: 10.1016/J.NEUCOM.2021.04.102]

Xu G, Liu Z and Loy C C. 2020. Computation-efficient knowledge distillation via uncertainty-aware mixup [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/2012.09413.pdf

Xu Z, Hsu Y C and Huang J W. 2018. Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/1709.00513.pdf

Yang C L, Xie L X, Su C and Yuille A L. 2019. Snapshot distillation: teacher-student optimization in one generation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 2854-2863 [DOI: 10.1109/CVPR.2019.00297]

Yao L W, Pi R J, Xu H, Zhang W, Li Z G and Zhang T. 2021. G-DetKD: towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/2108.07482.pdf

Yim J, Joo D, Bae J and Kim J. 2017. A gift from knowledge distillation: fast optimization, network minimization and transfer learning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 7130-7138 [DOI: 10.1109/CVPR.2017.754]

You S, Xu C, Xu C and Tao D C. 2017. Learning from multiple teacher networks//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada: Association for Computing Machinery: 1285-1294 [DOI: DOI: 10.1145/3097983.3098135]

Yue K Y, Deng J F and Zhou F. 2020. Matching guided distillation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 312-328 [DOI: 10.1007/978-3-030-58555-6_19]

Zagoruyko S and Komodakis N. 2017. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/1612.03928.pdf

Zhang L F, Song J B, Gao A N, Chen J W, Bao C L and Ma K S. 2019. Be your own teacher: improve the performance of convolutional neural networks via self-distillation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 3712-3721 [DOI: 10.1109/ICCV.2019.00381]

Zhang L F, Shi Y K, Shi Z Q, Ma K S and Bao C L. 2020a. Task-oriented feature distillation//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 14759-14771

Zhang Y, Xiang T, Hospedales T M and Lu H C. 2018. Deep mutual learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4320-4328 [DOI: 10.1109/CVPR.2018.00454]

Zhang Y C, Lan Z H, Dai Y C, Zeng F G, Bai Y, Chang J and Wei Y C. 2020b. Prime-aware adaptive distillation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 658-674 [DOI: 10.1007/978-3-030-58529-7_39]

Zhao H R, Sun X, Dong J Y, Chen C R and Dong Z H. 2019. Highlight every step: knowledge distillation via collaborative teaching [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/1907.09643.pdf

Zhi Z, Ning G H and He Z H. 2017. Knowledge projection for effective design of thinner and faster deep neural networks [EB/OL]. [2022-03-13]. https://arxiv.org/pdf/1710.09505.pdf

Zhou G R, Fan Y, Cui R P, Bian W J, Zhu X Q and Gai K. 2018. Rocket launching: a universal and efficient framework for training well-performing light net. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1): 4580-4587 [DOI: 10.1609/aaai.v32i1.11601]

Zhu J G, Tang S X, Chen D P, Yu S J, Liu Y K, Rong M Z, Yang A J and Wang X H. 2021. Complementary relation contrastive distillation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 9256-9265 [DOI: 10.1109/CVPR46437.2021.00914]

作者简介

司兆峰,男,硕士研究生,主要研究方向为神经网络压缩。E-mail: sizhaofeng19@mails.ucas.edu.cn

齐洪钢,通信作者,男,教授,主要研究方向为视频编解码和计算机视觉。E-mail: hgqi@ucas.ac.cn